DOCUMENT RESUME

ED 468 960 TM 034 448

AUTHOR	Lee, Yong-Won
TITLE	Investigating the Graded Relationships among Score Categories in Passage-Based Testlets: Application of the Nominal Response Model.
PUB DATE	2001-08-03
NOTE	35p.; Paper presented at the International Conference of the Korea English Education Society (1st, Chungbuk, Korea, August 2-3, 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	*English (Second Language); Foreign Countries; *High School Students; High Schools; Reading Comprehension; *Reading Tests; *Scores
IDENTIFIERS	Constraints; Graded Response Model; *Nominal Response Model; South Korea; *Testlets

ABSTRACT

The main purpose of this study was to examine the structure of score categories in passage-based testlets in an English as a Foreign Language (EFL) reading comprehension test using the nominal response model (NRM). The computer program MULTILOG (D. Thissen, 1991) was used to fit the full and 7 constrained versions of NRM to testlet scores of 1,857 Korean high school students from an EFL reading comprehension test made of 10 passage-based testlets. The results of the analysis have revealed that it would be reasonable to constrain the category-to-category changes in discrimination and distribution parameters to be linear and quadratic plus linear for the 10 testlets, respectively. Results also show that it would be appropriate to assume that the estimated threshold difficulty parameters increase linearly from the lower to higher steps. It is concluded that the score categories within each passage-based testlet could be practically regarded as graded responses. (Contains 3 figures, 3 tables, and 30 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 468 960

Investigating the Graded Relationships among Score Categories in Passage-Based Testlets: Application of the Nominal Response Model

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Y-W. Lee

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Yong-Won Lee, Ph. D.

Center for Validity Research

Division of Research and Development

Educational Testing Service (ETS)

Princeton, NJ 08541, USA

Phone: (609)-734-1348

Fax: (609)-734-1755

Email: ylee@ets.org

Paper included in the proceedings of the inaugural conference of Korea English Education Society (KEES) held at Korea National University of Education (KNUE) in Republic of Korea on August 3, 2001.

TM034448

BEST COPY AVAILABLE

Abstract

The main purpose of the study reported in this article is to examine the structure of score categories in *passage-based testlests* in an EFL (English as a Foreign Language) reading comprehension test by using the *nominal response model* (NRM). A computer program MULTILOG (Thissen, 1991) was used to fit the full and 7 constrained versions of NRM to testlet scores of 1,857 Korean high school students from an EFL reading comprehension test made up of 10 passage-based testlets. The results of the analysis have revealed that: (1) it would be reasonable to constrain the category-to-category changes in *discrimination* (a_{ik}) and *distribution* (c_{ik}) parameters to be linear and quadratic plus linear for the 10 testlets, respectively; (2) it would be appropriate to assume that the estimated *threshold difficulty* (b_{ij}) parameters increase linearly from the lower to higher steps. It is concluded that the score categories within each passage-based testlet could be practically regarded as graded responses.

Introduction

In EFL (English as a Foreign Language) reading comprehension tests, a series of related items are created based on a common reading passage. Such a cluster of items sharing stimulus material is often called a *testlet* in the assessment literature (Wainer & Kiely, 1987). When a test made up of *passage-based testlets* is analyzed through item response theory (IRT), *local item dependence* (LID) in passage-based testlets would pose a serious challenge for test analysis. The sharing of a passage among items is very likely to lead to the violation of the "local item independence" (LII) assumption required for IRT (Lee, *in press*; Schedl, Gordon, Carey & Tang, 1996; Thissen, Steinberg & Mooney, 1989, 1989). When IRT models assuming LII are applied to the tests that include sets of *locally-dependent* items, inaccurate estimation of item parameters, examinee abilities, and item and test information can result (Ferrara, Huynh & Michaels, 1999; Wainer & Thissen, 1996; Yen, 1993).

In the IRT framework, one possible approach to solving the problem of LID is to treat a set of related items based on the same passage as a polytomous super-item that yields more than two response categories and fit the polytomous IRT models to the test data (Cook, Dodd & Fitzpatrick, 1999; Thissen *et al.*, 1989; Wilson, 1988; Wilson & Adams, 1995). By using *polytomously-scored* testlets as units of analysis in IRT analysis, it would be necessary only to assume LII among the testlets, but not among items within passages (Rosenbaum, 1988; Wainer & Kiely, 1987). Another promising alternative approach might be to group together items that are *locally-dependent* and treat them as a cluster of items sharing an additional item parameter and fit the *testlet response model* to the data (Wainer & Wang, 2000). Nevertheless, the first approach is the focus of research

in this study for the following two reasons: (1) the estimation software for the *testlet response model* is not readily available yet for general users; (2) the current version of the software is not able to handle data sets with missing cases satisfactorily.

The main purpose of the study reported in this article was, therefore, to examine the structure of score categories in passage-based testlests in an EFL (English as a Foreign Language) reading comprehension test by using Bock's (1972) *nominal response model* (NRM). In this paper, the author (1) provides an overview of polytomous IRT models applicable for testlet-based tests, with a special focus on NRM; and (2) reports the results of fitting several constrained and unconstrained versions of the NRMs and two other polytomous IRT models to data from a 10-testlet-based EFL reading comprehension test using the computer program, MULTILOG (Thissen, 1991).

IRT Models for Passage-Based Testlets

Local Item Dependence (LID) in Passage-Based Testlets

Local item dependence (LID) can be defined as "lack of local independence between items" in the framework of IRT (Henning, 1989; Lord & Novick, 1968; Yen, 1993). In IRT, the assumption of *local item independence* (LII) stipulates that responses to two different items should be statistically independent of each other for examinees at a fixed ability/trait level (Hambleton & Swaminathan, 1985; Lord & Novick, 1968). In a weaker version of the assumption, LII could be defined technically as a zero correlation between the two items for individuals, after the impact of measurement construct (θ) is partialled out from the item scores (Henning, 1989; Weiss & Yoes, 1991; Yen, 1993). Thus, the two items are regarded as *locally-dependent*, if there is still a significant level of

a correlation remaining between the residuals of the two items, after the removal of the impact of the measurement construct from the item scores.

In reading comprehension tests, most of the items sharing a passage are basically designed to assess examinees' comprehension of the related informational elements in the same text. Examinees' performance on one item may be thus more highly correlated to their performance on another item in the same passage than on any item in a different passage. In an IRT sense, however, not all of the excessive interitem covariation within the passage may contribute to "a common latent trait" (θ) underlying a whole collection of items across different texts in the test (Such a common latent trait to be modeled in IRT might be conceptualized as a "general reading comprehension proficiency" generalizable across a particular text in this context). The extra "passage-specific" item covariation within passages, which is unaccounted for by this common latent trait, is very likely to lead to "correlated residuals" or LID. It may also indicate that there is a secondary test dimension (s) (*e.g.*, background knowledge about the passage, passage topic familiarity) tapped by these items, but unaccounted for by the IRT *theta* (θ).

In principle, if the content of one item provides clues to the answer of another item in a test, LID occurs (Hambleton & Swaminathan, 1985). In EFL reading comprehension tests, overlapped information between the options of one item and the stem of another item within a passage (Lee, *in press*) could become a potential source of LID for some items as well as the physical layout of the test booklet (Muraki & Lee, 2001). Yen (1993) has also provided a list of potential causes of LID including the sharing of a common passage, content, knowledge, item chaining, speededness, fatigue, practice effects, item or response format, and so forth. Chen and Thissen (1997) have

made a distinction between "the underlying LID" stemming from the contents of items and "the surface LID" often found in a series of not-reached items at the end of the test.

NRM as an Exploratory Tool for Passage-Based Testlets

Polytomous IRT and a testlet approach. One possible solution to LID in testlet-based tests is to treat a set of related items based on the same passage as a single polytomous item with more than two response categories and fit the polytomous IRT models to the test data (Cook *et al*, 1999; Thissen *et al.*, 1989; Wilson, 1988; Wilson & Adams, 1995). Examinees' responses to a testlet are scored polytomously, such as 0, 1, 2, ..., *m* depending on how many questions the examinees get right out of *m* questions in the testlet. Such an approach is sometimes called *a testlet approach* in the literature (Lee, 1999; Muraki & Lee, 2001; Wainer & Lewis, 1990). The *testlet approach* has been advocated on the grounds that the impact of local dependence among items within the testlet could be removed by treating them as a single item, whereas the local independence between the testlets might be still maintained (Rosenbaum, 1988; Wainer & Kiely, 1987).

Included among the polytomous IRT for testlets are Bock's (1972) nominal response model (NRM), Samejima's (1969) graded response model (GRM), Masters' (1982) partial response model (PCM), and Muraki' (1992) generalized partial credit model (GPCM). When score categories within testlets are found to be ordered as a matter of empirical fact, the last four polytomous IRT models for graded item responses could be fitted to the testlet data to estimate the testlet and ability parameters. When the relative ordering of score categories within the testlet with respect to the construct being measured is not clearly known *a priori*, however, these score categories can be first

treated as *nominal* (or at most semi-ordered) categories and the structure of the testlet scores can be investigated through the IRT model-data fitting process (Steinberg & Thissen, 1996). Under such circumstances, NRM could be an ideal exploratory tool for investigating the existence of graded relationships among the score categories with testlets as a first step of testlet-based analysis.

Nominal Response Model (NRM). Bock's (1972) nominal response model (NRM) assumes that item alternatives represent responses measured at a nominal level of measurement (*i.e.*, unordered response categories). NRM provides a direct expression for obtaining the probability of an examinee with ability (θ) responding in the *k*th category of item *i* as:

$$P_{ik}(\theta) = \frac{\exp(c_{ik} + a_{ik}\theta)}{\sum_{h=1}^{m_i} \exp(c_{ik} + a_{ik}\theta)},$$

where a_{ik} is the slope (discrimination) parameter and c_{ik} is the intercept (distribution) parameter of the nonlinear response function associated with the *k*th category of item *i*. In NRM, each category's ability to discriminate examinees is captured by the category's individual discrimination parameter, a_{ik} (De Ayala, 1993). A category with a larger a_{ik} reflects a response pattern, in which, as one progresses along the lower to higher ability continuum, there is a corresponding increase in the number of persons who answered the item in that category. On the other hand, larger values of c_{ik} are associated with categories with higher frequencies and are intended to capture the smoothness in the distribution of response proportions for score categories (De Ayala, 1993; Thissen *et al.*, 1989). To

identify the model, the following constraints are imposed on the $a_{ik}s$ and $c_{ik}s$:

$$\sum_{k=0}^{m_i} a_{ik} = \sum_{k=0}^{m_i} c_{ik} = 0.$$

The parameters for *difficulties of transitions* between adjacent score categories (sometimes called *threshold* or *step difficulty* parameters) can be computed as follows (De Ayala, 1993): $b_{ij} = c_{i(k-1)} - c_{ik} / a_{ik} - a_{i(k-1)}$,where b_{ij} is an item parameter associated with a difficulty for the *j*th threshold (*i.e.*, a transition between the *(k-1)*th and the *k*th category) of item *i*. When there are *k* score categories, there would be *(k-1)* thresholds difficulty parameters (*e.g.*, four possible b_{ij} parameters for 5 score categories). When an item has only two options (*i.e.*, correct or incorrect), NRM reduces to the two-parameter logistic (2PL) model, in which the item difficulty parameter (b_i) is equal to a value of $-c_i/a_i$ (Dodd, De Ayala & Koch, 1995).

Research Questions

In exploring the suitability of NRM and other polytomous IRT models in analyzing the testlet scores, the following four research questions were in mind:

1. Would it be reasonable to assume the category-to-category changes in *discrimination* ($a_{ik}s$) parameters to be linear for each testlet in an EFL reading comprehension test made up of *passage-based testlets*?
2. Would it be reasonable to assume the category-to-category changes in *distribution* ($c_{ik}s$) parameters to be quadratic plus linear for all the testlets used in the test?
3. Do the estimated *threshold difficulty* (b_{ij}) parameters increase linearly from the lower to higher thresholds for all the testlets used in the test?
4. Would polytomous IRT models for graded responses (e.g., PCM, GRM) be applicable for the test used in this study?

Method

Sample

Participants were 1,857 Korean EFL (*English as a Foreign Language*) high school students enrolled in 5 different high schools in the cities of Seoul and Incheon, South Korea during Fall 1997. There were 613 tenth graders and 1,244 eleventh graders. Of 1875 examinees, 1089 were male examinees and 768, female. The number of participants at each school ranged from 304 to 413.

Instrument

A 10-passage and 40-item multiple-choice reading comprehension test was developed for Korean high school students by the author. Two passages selected from each of the five subject areas of humanities, social sciences, natural sciences, engineering, and arts/sports. Four multiple-choice (MC) items representing 4 item types (*i..e.*, main topic inference, recall of details, inference of details, and prediction of content) were constructed for each passage. In previous research (Lee, *in press*), passage-related LID was confirmed across all 10 passages used in this study through Q_3 analyses.

The *standardized item alpha* (α) based on 40 items was .89, whereas the *passage alpha* (α) based on 10 passage-testlets was .87. The marginal reliability estimated by fitting the 3-parameter IRT model to the 40 item test was .90[1]. The scale mean and standard deviation for the total score were approximately 20.4 and 8.58, respectively (See Lee (*in press*) for more details about the instrument and data).

---

[1] In IRT, the precision of measurement is assumed to vary as a function of examinees' ability (θ). Nevertheless, a marginal reliability can be computed to provide an IRT analog of internal consistency Reliability (Thissen, 2000). The marginal reliability is based on marginal error averaged over all ability (θ) ranges (Green, Bock, Humphreys, Linn & Reckase, 1984).

Data Analysis

A computer program MULTILOG (Thissen, 1991) was used to fit the full NRM and 7 constrained NRMs with various combinations of parameter constraints for the *discrimination* (a_{ik}) and *distribution* (c_{ik}) parameters for the 10 testlets. The marginal maximum likelihood (MML) estimation method was employed in MULTILOG to estimate the testlet parameters, with the maximal number of EM (expectation-maximization) cycle was set at 100. The convergence criteria for the M-step and EM cycle were set at 0.001. Since there were 5 possible categories (*i.e.*, 0, 1, 2, 3, 4) in each passage-based testlet in this study, both a_{ik}s and c_{ik}s had to be estimated for each of these 5 score category. However, only 4 *threshold difficulty* (b_{ij}) parameters had to be computed for each testlet by using the estimated a_{ik} and c_{ik} parameters, as defined by Dodd, De Ayala, and Koch (1995).

In MULTILOG, polynomial contrasts were used to represent the category-to-category changes in a_{ik}s and c_{ik}s for the 10 testlets. Because there were 5 score categories in each testlet, all of the linear [$\alpha_i(1)$], quadratic [$\alpha_i(2)$], cubic [$\alpha_i(3)$], and quartic [$\alpha_i(4)$] polynomial terms were possible for the a_{ik}s for each testlet, with the same four levels of polynomial contrasts [*i.e.*, $\gamma_i(1)$, $\gamma_i(2)$, $\gamma_i(3)$, $\gamma_i(4)$] possible for the c_{ik}s as well. In the full NRM, these 8 polynomial contrasts were all of the "free parameters" that had to be estimated for each testlet (*i.e.*, a total of 80 free parameters for the 10-testlet test). In the first phase of creating constrained NRMs, the $\alpha_i(4)$, $\alpha_i(3)$, and $\alpha_i(2)$ terms were constrained to be zero one after the other for all 10 testlets at a time, with all of the $\gamma_i(1)$, $\gamma_i(2)$, $\gamma_i(3)$, and $\gamma_i(4)$ allowed for the c_{ik}s. In the next phase, the $\gamma_i(4)$, $\gamma_i(3)$, and $\gamma_i(2)$

terms were constrained to be zero again in a stepwise fashion, with only the linear term [$\alpha_i(1)$] allowed for the a_{ik}s.

The analysis results of model-data fit and marginal reliability from the full NRM were first compared to those from the constrained versions of the NRMs. Model-data fit statistics and marginal reliability coefficients from the full NRM were also compared to those from the graded response model (GRM) and an MML version of the partial credit model (PCM) used in this study. Moreover, $\alpha_i(p)$s and $\gamma_i(p)$s estimated from the full NRM were plotted for all of the 10 testlets, with the a_{ik}s and c_{ik}s plotted for the 10 passages as well. The *threshold difficulty* (b_{ik}) parameters for the 10 passage-based testlets from the full NRM, PCM, and GRM were plotted and compared.

Results

The results of model-data fit analyses for NRM, PCM, and GRM and the plots of the $\alpha_i(p)$s, a_{ik}s, $\gamma_i(p)$s, c_{ik}s, and b_{ij}s are presented in the following order: (1) comparison of the model-data fit statistics from the full and constrained NRMs; (2) the plots of the estimated $\alpha_i(p)$s, a_{ik}s, $\gamma_i(p)$s, and c_{ik}s from the full NRM; (3) comparison of model-data fit statistics and the plots of the estimated b_{ij}s from the full NRM, GRM, and PCM.

Model-Data Fit Statistics of the Full and Constrained NRMs

Table 1 shows the model-data fit statistics and the marginal reliability coefficients for the full NRM and the 7 constrained NRMs fitted to the testlet score data prepared for this study. First, it was observed that the $\alpha_i(p)$s can be constrained to be linear for all of the 10 testlets without exacerbating the model-data fit statistics seriously, compared to the full NRM. Among the 8 comparison NRMs, it turned out that the full NRM apparently had the best model-data fit, with the smallest negative 2loglikelihood statistics (17,582.2)

and the largest marginal reliability coefficient (0.85)[2]. As the $\alpha_i(4)$, $\alpha_i(3)$, and $\alpha_i(2)$ terms were constrained to be zero one after the other for all of the 10 testlets at a time, the negative 2loglikelihood increased by 14.60 ($p>0.1$), 31.80 ($p<0.001$), and 79.20 ($p<0.001$), respectively, but the marginal reliability coefficient remained at .85 for the 3 constrained NRMs (See Table 1). When the a_{ik}s were also further constrained to be equal across 10 testlets (as in PCM) in addition to the $\alpha_i(p)$ being constrained to be linear, the marginal reliability was lowered to .84 and the negative 2loglikelihood increased by 77.20 ($p<0.001$).

Insert Table 1 about here.

Second, it seemed reasonable to constrain the $\gamma_i(p)$ to be quadratic plus linear for most of the testlets without damaging the model-data fit statistics seriously, but may be hard to assume that the $\gamma_i(p)$s are linear for each testlet. As the quartic [$\gamma_i(4)$] and cubic [$\gamma_i(3)$] polynomial terms for the c_{ik}s were constrained to be zero one after the other for all 10 testlets, with only the $\alpha_i(1)$ retained for the a_{ik}s in the model, the negative 2loglikelihood value increased by 55.2 (df=10, p<.001) and 113.2 (df=10, p<.001), respectively, but the marginal reliabilities remained at .85 and dropped to .84, respectively. When the quadratic [$\gamma_i(2)$] term was further constrained to be zero, however, the negative 2loglikelihood statistics increased far more drastically (*i.e.*, increase in χ^2 = 3317.8, df=10, p<.001) and the marginal reliability dropped steeply down to .74.

---

[2] Because NRM models 4 and 6 in Table 1 are identical in substance, the two models were counted only once toward the total number of the comparison models.

Estimated Polynomial Contrasts and Item Parameters from the full NRM.

Figures 1-a and 1-b show the estimated $\alpha_i(p)$s and a_{ik}s from the full NRM, before any testlet parameter constraints were imposed on the model. First, as shown in Figure 1-a, the $\alpha_i(2)$s, $\alpha_i(3)$s, and $\alpha_i(4)$s for the 10 testlets were narrowly clustered around 0, whereas the $\alpha_i(1)$s were significantly larger than 0 for all of the 10 testelts, which ranged from 0.52 to 0.93, with the mean of 0.75 and the standard deviation of 0.13. Second, these results seemed to be consistent with the pattern of changes in a_{ik}s over 5 score categories displayed in Figure 1-b. It seems that the category-to-category change in a_{ik}s could be well represented by linear lines for most testlets. Even though a few testlets (*e.g.*, P02, P05) appeared to have some quadratic (curvilinear) as well as linear effects, the lines connecting a_{ik}s over 5 score categories for even these two testlets did not deviate seriously from the straight lines, as shown in Figure 1-b.

Insert Figures 1-a and 1-b about here.

Figures 2-a and 2-b show graphically the estimated $\gamma_i(p)$s and $c_{ik}s$ from the full NRM. It is seen in Figure 2-a that the $\gamma_i(3)$s and $\gamma_i(4)$s for the 10 testlets are narrowly clustered near 0 with the standard deviations of 0.04 and 0.01, respectively, whereas the $\gamma_i(1)$s were more widely clustered around zero, ranging from -0.13 to 0.23, with the larger standard deviation of 0.11. Interestingly enough, it was actually the $\gamma_i(2)$ effects that were the largest in magnitude among the four levels of polynomial contrasts used for the $c_{ik}s$.

The $\gamma_i(2)$ coefficients ranged from –0.53 to –0.34 for the 10 testlets, with the mean of –0.42 and the standard deviation of 0.07.

Such dominance of quadratic (curvilinear) effects for c_{ik}s in the 10 testlets was also confirmed in the plot of c_{ik}s over score categories. As shown in Figure 2-b, the lines connecting c_{ik}s over 5 score categories were shaped much like inverse "U" curves for most of the 10 testlets. It was also revealed in Figure 2-b that the frequencies for middle score categories, such as 2 and 3, were comparatively high, with the extreme score categories (*i.e.*, 0, 4) having lower frequencies. Although the two trend lines of c_{ik}s for Testlets 3 (P03) and 5 (P05) seemed to be a little more negatively skewed than those for the rest of the testlets, they did not deviate seriously from the "inverse U-shaped curves" either.

Insert Figures 2-a and 2-b about here.

Model-Data Fit and estimated b_{ik}s from Full NRM, GRM, and PCM

Model-Data Fit Analyses. Table 2 displays the model-data fit statistics and the marginal reliability coefficients from the full NRM, GRM, and PRM for comparison. Among these 3 comparison models, the full NRM exhibited the best model-data fit, with the smallest negative 2loglikelihood value (17582.2), but GRM and PCM showed the somewhat worse model-data fit than the full NRM, with the negative 2log likelihood values of 17768.5 and 17,785.0, respectively. Compared to the full NRM, the total number of "free parameters" to be estimated in GRM decreased from 80 to 50 (by 30), the negative 2log likelihood statistics increased by 186.3, but the marginal reliability also

remained at .85. The model-data fit for GRM did not seem to be seriously bad, compared to its NRM equivalent (model 4 in Table 1). The difference in the negative 2loglikelihood statistics between the full NRM and the constrained NRM with only the linear polynomial terms [$\alpha_i(1)$] for a_{ik}s (Model 4) were about 125.6, whereas that between the full NRM and GRM was about 186.3.

On the other hand, it was found that the same values of the model-data fit statistics (17,785.0) were obtained for both the PCM and one of the constrained NRM (Model 5 in Table 1) that had the equal a_{ik}s assumed across 10 testlets in addition to the $\alpha_i(p)$ being constrained to be linear. The only difference between the two models was that the "triangle T-matrices" were used to represent the category-to-category change in $c_{ik}s$ in PCM, whereas the polynomial contrasts were used for Model 5. As expected, the PCM showed the slightly worse model-data fit statistics than GRM, and the marginal reliability was slightly lowered to .84. Since the full NRM and GRM were not the nested models in a strict sense, it was not appropriate to evaluate the statistical significance of an increase in the negative 2loglikelihood statistic against a reduction in the number of freedom.

Insert Table 2 about here.

Estimated $b_{ij}s$ from the full NRM, GRM, and PCM. Table 3 shows the estimated $b_{ij}s$ for each threshold level for the 10 testlets from the full NRM, GRM, and PCM. Figures 3-a, 3-b, and 3-c graphically display the patterns of changes in $b_{ij}s$ over 4 thresholds that were estimated separately from the three models for the 10 testlets. First, as shown in Figure 3-a and Table 3, the *threshold difficulty* parameters (b_{ij}s) from the full

NRM increased steadily from the lower to higher score categories. Moreover, the lines connecting the b_{ij} values from the lower to higher thresholds were close to straight lines in shape, and all of the four thresholds were ordered as expected in difficulty for most of the testlets. However, the b_{ij} trend lines for some testlets (e.g., P02) deviated significantly from the straight lines, especially at the lowest threshold (*i.e.*, a transition between 0 and 1). It seems that the difficulty parameter for the first threshold was much lower than expected for a straight line (See Figure 3-a). Somewhat smaller, but noticeable deviations of similar type were also observed for the lowest threshold for Testlets 8 and 9. Nevertheless, no adjacent thresholds were reversed in difficulty in any of the 10 testlets.

Insert Table 3 and Figures 3-a, 3-b, and 3-c about here.

Second, the *threshold difficulty* (b_{ij}s) parameter both from the GRM and PCM increased steadily in a linear fashion from the lowest through highest thresholds (See Figures 3-b and 3-c). The lines connecting the b_{ij}s over the 4 thresholds seemed to be quite similar to straight lines for all of the 10 testlets in both GRM and PCM. Besides, all of the four thresholds were nicely ordered as expected in difficulty for all of the testlets. Especially, the irregularities of b_{ij}s observed at the lowest threshold for some testlets in the full GRM disappeared in GRM and PCM.

When the three different b_{ij}s from the full NRM, GRM, and PCM are compared at each threshold, the three different b_{ij}s were not much different from one another in most testlets (See Table 3). As mentioned above, one of the outstanding exceptions is the b_{ij}s at

the lowest threshold for Testlet 2. The b_{ij} value (–4.56) from the full NRM was unusually lower than those from GRM (–2.48) and PCM (–2.14) at this threshold.

Summary and Discussion

The main objective of the study reported in this article was to investigate the structure of score categories in passage-based testlests by using the nominal response model (NRM), with a view to exploring the applicability of existing polytomous IRT models for graded responses in an EFL reading comprehension test. The major findings of the study are summarized and discussed in terms of (1) linear changes in *discrimination* parameters (a_{ik}s), (2) curvilinear plus linear changes in *distribution* parameters (c_{ik}s), and (3) graded structure of score categories within testlets.

Linear Changes in Discrimination Parameters

The plotting of the estimated $\alpha_i(p)$s and a_{ik}s from the full NRM turned out to be a powerful tool for evaluating the results of the model-data fit analyses of constrained and unconstrained versions of NRMs with a_{ik}s constraints. First, the model-data fit analyses of the full and constrained NRMs have revealed that the category-to-category changes in a_{ik}s within the 10 testlets could be modeled by the linear polynomial terms ($\alpha_i(1)$) alone without sacrificing the model-data fit seriously in the test. There were some significant increases in the negative 2loglikelihood ($p<0.001$), when the cubic and quadratic terms were constrained to be zero in a stepwise fashion. However, such significant increases in the χ^2 statistics might not be taken too seriously, if we take into consideration the sample size and the reduced number of free parameters to be estimated in both models.

Second, such results of the model-data fit analyses become much clearer, when the $\alpha_i(p)$s and a_{ik}s from the full NRM were plotted for the 10 passages. Although all of

the quartic, cubic, quadratic, and linear terms were allowed for the a_{ik}s in the full NRM model, it can be clearly seen in the plot of the $\alpha_i(p)$s that the values of estimated quadratic [$\alpha_i(2)$], cubic [$\alpha_i(3)$], and quartic [$\alpha_i(4)$] polynomial contrast coefficients were virtually negligible, whereas the linear contrast coefficients [$\alpha_i(1)$] were significantly larger than zero.

Last, such findings were also supported by the pattern of changes in a_{ik}s over 5 score categories displayed in the plot of a_{ik}s. It seems obvious that the category-to-category changes in a_{ik}s might be well represented by linear lines for most testlets. This may explain why the negative 2loglikelihood statistics did not increase drastically, when the quartic, cubic, and quadratic polynomial contrast coefficients for the a_{ik}s were constrained to be 0 in a stepwise fashion for the 10 testlets. As a matter of fact, there were a few testlets (*e.g.*, P02, P05) that appeared to have some quadratic (curvilinear) as well as linear effects to some extent. Nevertheless, these two testlets did not deviate seriously from the straight lines either. All these taken together, it would seem reasonable in a practical sense to model the category-to-category changes in the a_{ik}s to be linear for the 10 testlets.

Curvilinear Plus Linear Changes in Distribution Parameters

The results of the model-data fit analyses of the full NRM and the constrained NRMs with parameter constraints on c_{ik}s were again supported by the plotted patterns of the estimated $\gamma_i(p)$s and c_{ik}s from the full NRM. First, it was also observed that the category-to-category change in c_{ik}s could be modeled by the quadratic and linear terms without damaging the model-data fit much. As the quartic [$\gamma_i(4)$] and cubic [$\gamma_i(3)$]

polynomial terms for the c_{ik}s were constrained to be zero in a stepwise fashion for all 10 testlets, with only the $\alpha_i(1)$ retained for the a_{ik}s in the model, the negative 2loglikelihood value increased significantly ($p<.001$) in both cases, but the marginal reliability remained at .85 and dropped to .84, respectively. When the quadratic [$\gamma_i(2)$] term was further constrained to be zero, however, the negative 2loglikelihood statistics increased drastically and the marginal reliability dropped sharply down to .74 (by .10), suggesting that it might be unreasonable to assume that the $\gamma_i(p)$s are linear for each testlet.

Second, the results of the model-data fit analyses were found to be consistent with the patterns of the $\gamma_i(p)$s and c_{ik}s from the full NRM plotted for the 10 passages. Interestingly, it was the $\gamma_i(2)$ effects that were the largest in magnitude among the four levels of polynomial contrasts used for the c_{ik}s, whereas the $\gamma_i(3)$ and $\gamma_i(4)$ values for the 10 testlets were negligibly small (virtually zero). Nevertheless, the $\gamma_i(1)$ values were more widely clustered around zero than the $\gamma_i(3)$ and $\gamma_i(4)$ values. Such patterns of the $\gamma_i(p)$s were reflected again in the plots of the c_{ik}s. The trend lines connecting c_{ik}s for 5 score categories were shaped more like an inverse "U" curve for all of the 10 testlets, which may suggest that the relationship among c_{ik}s for different score categories can be more adequately represented by curvilinear lines.

As previously mentioned, the c_{ik}s parameter is basically the intercept parameter of the nonlinear response functions associated with the *k*th category of testlet *i*. For this reason, it would be reasonable to see that the frequencies for middle score categories, such as 2, 3 and 4, are comparatively higher that those for extreme score categories (*i.e.*,

0, 4). This may also explain well why the -2loglikelihood statistics increased more drastically, when the quadratic terms for the c_{ik}s were constrained to be 0.

Graded Structure of Score Categories within Testlets

An additional attempt was made in this study to explore the structure of score categories within testlets and the applicability of polytomous IRT models for graded responses in the testlet-based test by comparing the model-data fit statistics of the full NRM, PCM, and GRM and the estimated b_{ij}s from these three models. First, both GRM and PRM produced significantly larger negative 2log likelihood statistics than the full NRM. The model-data fit statistics of GRM and PRM were not seriously worse than those of the full NRM, given the reduced numbers of "free parameters" in both models. In addition, the marginal reliabilities remained at .85 and .84 for GRM and PCM, respectively.

Second, the comparison of the estimated *threshold difficulty* (b_{ij}s) parameters from the full NRM, PCM, and GRM has revealed that it would be appropriate to treat 5 score categories in each of the 10 testlets as graded responses. In the full NRM, the b_{ij}s increased steadily as score categories moved from the lowest to highest categories. Moreover, the lines connecting the b_{ij}s estimates from the lower to higher thresholds were similar to straight lines in overall shapes. There were a few testlets (e.g., P02, P08, P09) in which the b_{ij}s trend lines deviated significantly from the straight lines. The difficulty parameter for the first threshold (b_{i1}) was much lower than expected for a straight line. Nevertheless, no adjacent thresholds were reversed in difficulty even in this testlet. It should be noted that the b_{ik}s are not required to be ordered in the full NRM and PCM, whereas they are in GRM.

Intriguingly enough, Testlets 2 happened to be the one that was previously flagged due to small quadratic effect identified for the a_{ik}s at the lower score categories (*i.e.*, 0, 1) in addition to the overall linear effect across categories. More specifically, the rates of the a_{ik} change between score categories of 0 and 1 was much slower in Testlet 2 than those in other testlets, which most possibly have caused an extremely low threshold difficulty parameter for the first threshold level (*i.e.*, a transition between 0 and 1) for the testlet. Testlet scores are basically sums of *dichotomously-scored* MC item scores in each passage. Thus, it may mean that it would be easy to obtain a score of 1 (getting one item correct) through guessing and thus a testlet score of 1 may represent almost equally low proficiency as a testlet score of 0 (*Thissen et al.*, 1986).

In a sense, NRM is an extension of the 2-parameter logistic model in which no item parameter exists to model examinees' guessing behavior. For this reason, such a pseudo-guessing effect might be partially reflected into the patterns of category-to-category changes in the a_{ik}s and the threshold-to-threshold changes in the b_{ik}s.

Limitations, Avenues for Future Research, and Conclusion

Limitations and Avenues for Future Research

First, each level of polynomial parameter constraints for a_{ik}s and c_{ik}s were imposed on all of the 10 testlets as a whole than on one testlet at a time in order to simplify the model-data fitting process. It might have been more informative to allow a combination of different levels of the polynomial contrasts to be imposed on different testlets, because there were variations among the 10 testlets in terms of the patterns of category-to-category change in a_{ik}s and c_{ik}s. For instance, quadratic plus linear polynomial a_{ik}constrasts could have been imposed for some testlets (e.g., P02), while only

the linear contrasts are allowed for the rest of the testlets. Such an approach might have made it possible to do more fine-grained comparisons among the constrained NRMs rather than imposing the same level of constraints for all of the 10 testlets as a whole.

Second, the focus of this study was on examining the structure of the score categories in the testlet scores and the appropriateness of using the polytomous IRT models for passage-based testlets. However, the final goal of reading assessment would be to estimate examinees' reading proficiencies or abilities in most cases. In that sense, the nature of this study is somewhat preliminary. A more important step that should come next might be to do ability (θ) estimation based on the polytomous IRT models selected in this study and examine the impact of model selection on the accuracy of examinee ability estimates in the testlet-based tests.

Third, the *testlet approach* has some disadvantages in the sense that individual items are not identifiable and information on individual items is lost, after individual item scores are aggregated into testlet scores. As mentioned previously, one promising alternative would be to use the *testlet response model* (Wainer & Wang, 2000; Wang, Wainer & Bradlow, 2000), which is an extension of the dichotomous logistic models (*e.g.*, 3PL model). In this approach, individual items are still identifiable, information on individual items is not lost, and LID effect is modeled by an additional item parameter shared among items within the testlet (Wainer & Wang, 2000). As previously explained, this approach was not employed in this study due to some technical constraints. Nonetheless, further comparisons among dichotomous and polytomous IRT models and the *testlet response model* would inform us of strengths of these different approaches in modeling LID appropriately in the context of reading comprehension assessment.

Conclusion

Score categories in all of the 10 testlets can be largely regarded as graded responses and the polytomous IRT models for graded responses could be applied to the reading comprehension test data used in this study. NRM has also proven to be a powerful, exploratory tool for examining the graded relationships among score categories in the passage-based testlets in the context of EFL reading assessment. Particularly, the plotting of the estimated polynomial contrasts and testlet category parameters from the full NRM turned out to be useful and informative in investigating the structure of score categories within passage-based testlets in the context of reading assessment.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are stored in two or more latent categories. *Psychometrika, 37*, 29-51.

Chen, W., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Educational and Behavioral Statistics, 22, 265-289.*

Cook, K. F., & Dodd, B. G., & Fitzpatrick, S. J. (1999). A comparison of three polytomous item response theory models in the context of testlet scoring. *Journal of Outcome measurement,3(1), 1-20.*

De Ayala, R. J. (1993). An introduction to polytomous item response models. *Measurement and Evaluation in Counseling and Development, 25*, 172-189.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5-22.

Ferrera, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Assessment, 36*, 119-140.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Henning, G. (1989). Meaning and implications of the principle of local independence. *Language Testing, 6*, 95-108.

Lee, Y. (March, 1999). *A testlet approach to ESL reading assessment*. Paper presented at the 33rd annual convention of Teachers of English to Speakers of Other Languages (TESOL) in New York.

Lee, Y. (in press). Examining passage-related local item dependence (LID) and measurement construct using Q_3 statistics in an EFL reading comprehension test. *Language Testing*.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E., & Lee. Y. (2001). *Detecting local item dependency in the TOEFL reading comprehension section: An application of the full-information item factor analysis*. ETS internal document. Princeton, NJ: ETS.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53*, 349-359.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded score. *Psychometrika Monograph Supplement* No. 17.

Schedl, M., Gordon, A., Carey, P., & Tang, K. L. (1996). *An analysis of the dimensionality of TOEFL reading comprehension items* (TOEFL research report 53). Princeton, NJ: ETS.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*, 81-97.

Thissen, D. (1991). *MULTILOG user's guide*. Chicago: Scientific Software.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Eds.), Computer-adaptive testing: A primer (2nd Ed.) (pp. 159-184). Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Mooney J. A. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26*, 247-260.

Wainer, H., & Lewis, C. (1990). Toward a psychometric for testlets. *Journal of Educational Measurement, 27*, 1-14.

Wainer, H. & Kiely G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H., &Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203-220.

Wang, X., Bradlow, E. T., & Wainer, H. (2000). *User's guide for SCORIGHT (version 1.0): A computer program for scoring tests built of testlets*. Princeton, NJ: ETS.

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. In R. K. Hambleton & J. N.Zaal (Eds.), *Advances in educational and psychological testing: Theory and Applications* (pp. 69-95). Norwell, MA: Kluwer-Academic.

Wilson, M. (1988). Detecting and interpreting local item dependence using a family of Rasch models. *Journal of Applied Psychological Measurement, 12*, 353-364.

Wilson, M., & Adams, R. (1995). Rasch models for item bundles. *Psychometrika, 60*, 181-198.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Author Note

This paper is a revised portion of the author's Ph. D. dissertation. Needless to say, the responsibility for any errors that remain are solely mine, and the ideas and opinions expressed in the paper do not represent in any way those of Educational Testing Service (ETS). Address for correspondence: Requests for reprints should be sent to Yong-Won Lee, Mail Stop 19-R Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA; email: ylee@ets.org.

Table 1

Model-Data Fit and Marginal Reliabilities for Full and Constrained NRMs (N=1857)

Nested NRM Models	Marginal Reliability	No. of Free Parameters	Negative 2loglikelihood	χ^2 Increase	Decrease in df	*p*
(Step 1: Stepwise α(p) constraint for a_{ik}s, with all of *γ(1), γ(2), γ(3),* and *γ(4)* allowed for c_{ik}s)						
1. Full NRM [No *α(p) & γ(p)* constraint]	.85	80	17582.2			
				14.60	10	>.1
2. *α(4)*=0	.85	70	17596.8			
				31.80	10	<.001
3. *α(3, 4)*=0	.85	60	17628.6			
				79.2	10	<.001
4. *α(2,3,4)*=0*	.84	50	17707.8			
				77.2	10	<.001
5. α(2,3,4)=0 & a_{ik} equal for all testlets**		41	17785.0			
(Step 2: Stepwise $\gamma_i(p)$ constraint for c_{ik}s, with only *α(1)* allowed for a_{ik}s)						
6. No *γ(p)* constraint*	.85	50	17707.8			
				55.2	10	<.001
7. *γ(4)*=0	.85	40	17763.0			
				113.2	10	<.001
8. *γ(3,4)*=0	.84	30	17876.2			
				3317.8	10	<.001
9. *γ(2,3,4)*=0	.74	20	21194.0			

Note: 1. The two asterisked (*) models (models 4 and 6) are the same constrained model in substance.

2. The double-asterisked (**)model is the same as the MML (marginal maximum likelihood) estimation version of Masters' (1982) partial credit model (Thissen, 1991) appearing in Table 2, except for the contrast coefficient for the c_{ik}s. However, it should be noted that the "triangle contrast coefficients" were used for the c_{ik}s in the PCM, whereas the "polynomial contrast coefficients" were used for both the a_{ik}s and c_{ik}s in the double-asterisked model (model 5).

Table 2
Model Fit Statistics for 3 Polytomous IRT Models (N=1857)

IRT Models	Marginal Reliability	No. of Free Parameters	Negative 2loglikelihood	χ^2 Difference	Decrease in df
Full NRM	.85	80	17582.2		
				186.3	30
GRM	.85	50	17768.5		
				16.5	9
PCM	.84	41	17785.0		

Note: The asterisked (*) model is the MML (marginal maximum likelihood) version of Masters' (1982) *partial credit model* (Thissen, 1991). In this model, the quadratic and higher terms [*i.e.*, $\alpha(2)$, $\alpha(3)$, $\alpha(4)$] for the polynomial contrasts are constrained to be zero for the a_{ik}s and the triangle matrices are used for the contrasts for the c_{ik}s in the model. In addition, the a_{ik}s were assumed to be equal across all the testlets in addition to the linear term [$\alpha(1)$] assumed for the a_{ik}s for the testlets.

Table 3

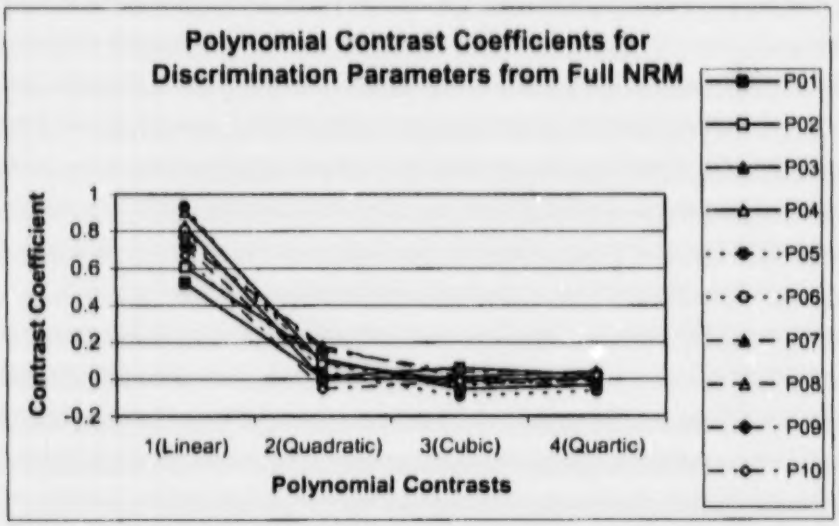
Estimated Threshold Difficulties (b_{ij}) for 10 Passages from the Full Nominal Response Model (NRM), the Graded Response Model (GRM), and the Partial Credit Model (PCM)

Testlet No.	Threshold 1 (0-1)			Threshold 2 (1-2)			Threshold 3 (2-3)			Threshold 4 (3-4)		
	NRM	GRM	PCM	NRM	GRM	PCM	NRM	GRM	PCM	NRM	GRM	PCM
P01	-1.55	-2.33	-1.59	-0.51	-0.71	-0.49	0.35	0.66	0.32	1.73	2.50	1.92
P02	-4.56	-2.48	-2.14	-0.59	-0.87	-0.47	0.03	0.33	0.05	1.58	1.91	1.54
P03	-1.13	-1.71	-1.08	-0.68	-0.81	-0.77	-0.23	0.08	-0.22	0.98	1.30	1.07
P04	-1.52	-1.62	-1.36	-0.40	-0.53	-0.35	0.21	0.41	0.19	1.36	1.65	1.54
P05	-1.21	-1.67	1.09	-1.21	-1.67	-1.09	-0.69	-0.80	-0.62	1.84	1.52	1.62
P06	-1.42	-1.52	-1.20	-0.06	-0.36	0.00	0.24	0.51	0.23	1.25	1.65	1.31
P07	-1.71	-2.02	-1.70	-0.48	-0.66	-0.43	0.25	0.45	0.26	1.56	1.83	1.51
P08	-2.81	-1.88	-1.70	-0.43	-0.64	-0.32	0.23	0.35	0.36	0.83	1.39	0.82
P09	-3.05	-2.29	-2.14	-0.91	-0.93	-0.82	0.23	0.29	0.31	1.07	1.59	1.27
P10	-1.92	-2.27	-1.86	-0.51	-0.71	-0.58	0.47	0.65	0.41	2.61	2.42	2.24

Figure 1

Estimated Polynomial Contrast Coefficients [$\alpha_i(p)s$] and Discrimination Parameters (a_{ik}s) for the 10 Passage-Based Testlets from the Full Nominal Response Model (NRM)

(a) Estimated polynomial contrast coefficients for a_{ik}s [$\alpha(p)$]



(b) Change of estimated discrimination parameters (a_{ik}s) over score categories

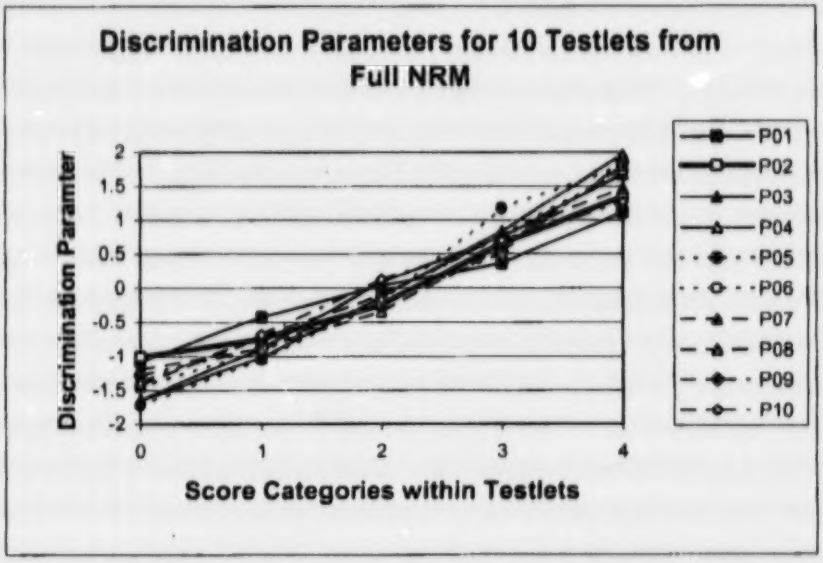
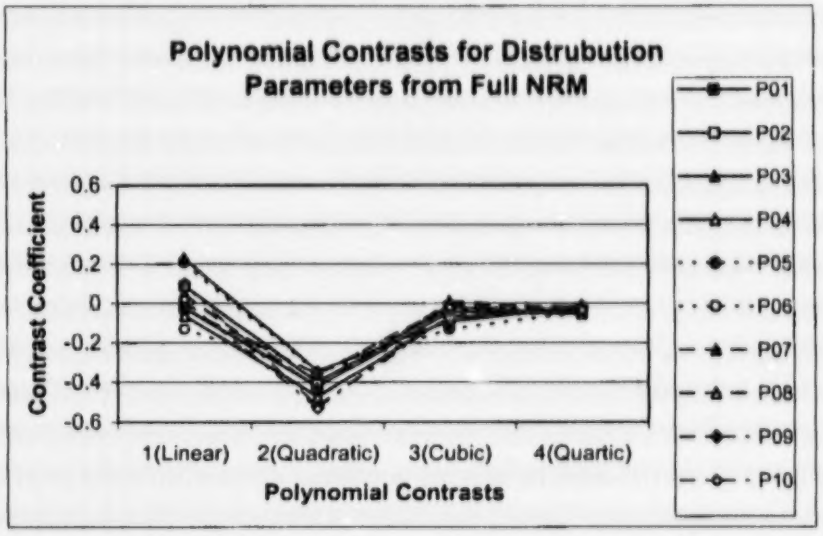


Figure 2

Change of a_{ik}s and c_{ik}s over 5 Score Categories Estimated for the 10 Passage-Based Testlets from the Full Nominal Response Model (NRM)

(b) Estimated polynomial contrast coefficients [$\gamma(p)$] for c_{ik}s



(b) Change of estimated distribution parameters (c_{ik}s) over score categories

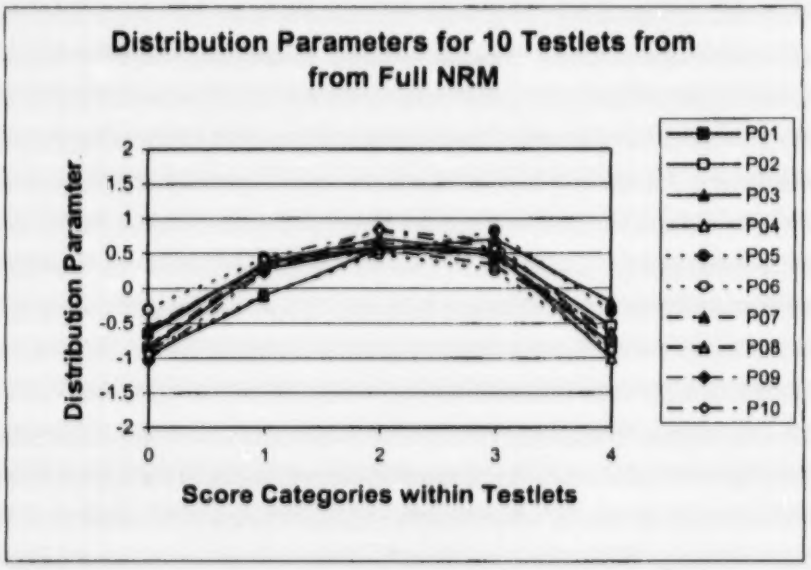
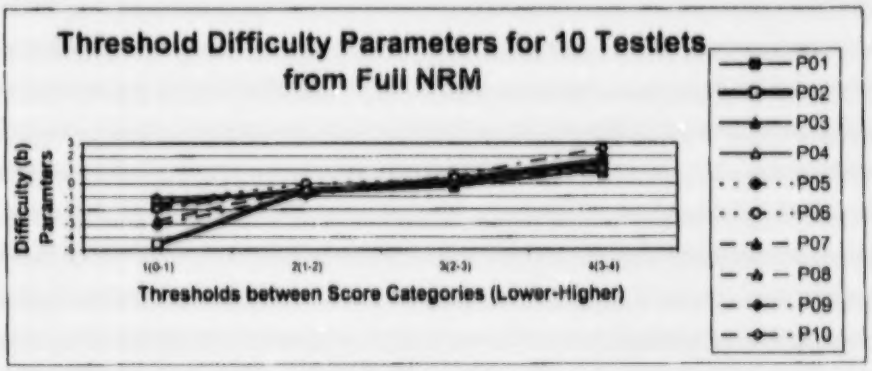


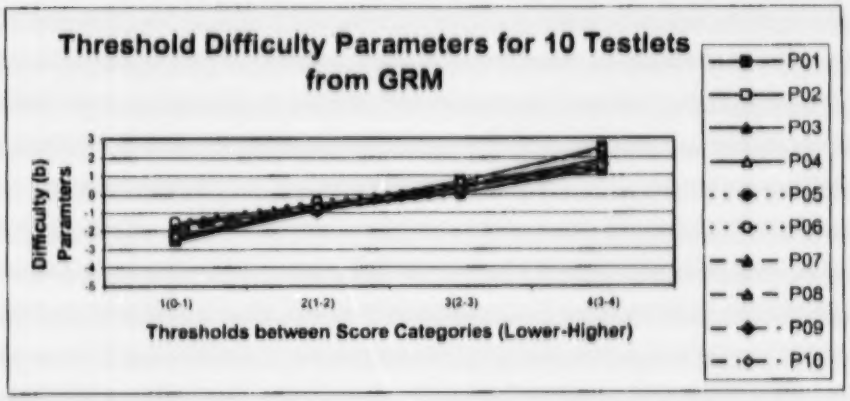
Figure 3

Estimated Threshold Difficulties for 10 Passages from the Full Nominal Response Model (NRM), the Graded Response Model (GRM), and the Partial Credit Model (PCM)

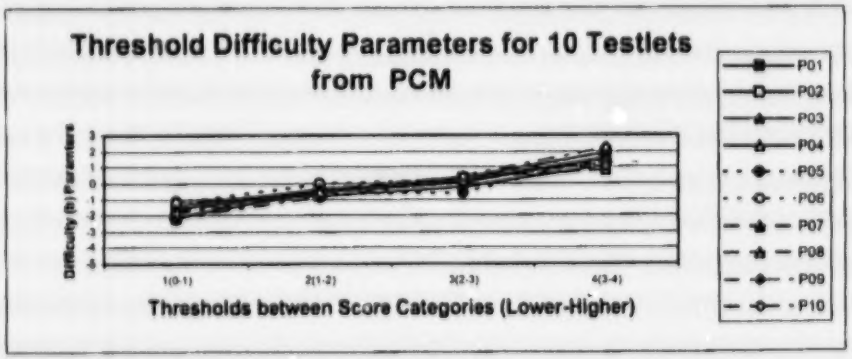
(a) Full NRM



(b) GRM



(c) PCM



TM034448

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Investigating the Graded Relationships among Score Categories in Passage-Based Testlets: Application of the Nominal Response Model	
Author(s): Yong-Won Lee	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ☐	Level 2A ☒	Level 2B ☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Yong-Won Lee		Printed Name/Position/Title: Associate Research Scientist	
Organization/Address: Educational Testing Service (ETS) Rosedale Road Princeton, NJ 08541		Telephone: 1-609-734-1382	FAX: 1-609-734-1755
		E-Mail Address: ylee@ets.org	Date: 9/12/2002

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org

EFF-088 (Rev. 2/2001)